Case study on XML processing Hive XML SerDe

Hive XML SerDe is an XML processing library based on Hive SerDe (serializer / deserializer) framework. It relies on XmlInputFormat from Apache Mahout Project to shred the input file into XML fragments based on specific start and end tags.

Data: The dataset ebay.xml consists of the following columns:

seller_name

seller_rating

bidder_name

location

bid_history

item_info

Hive does not have any built-in support for XML data processing, so we need to use the XML SerDe that have been made available by open source developers.

Download the latest version of *hivexmlserde.jar* from here and copy it to your /lib folder.

1. Add the XML SerDe jar into the hive shell

ADD JAR /home/cloudera/hivexmlserde-1.0.5.3.jar /usr/lib/hive/lib;

2. Create a table for the given XML data

CREATE TABLE ebay_listing(seller_name STRING,

seller_rating BIGINT, bidder_name STRING,

location STRING, bid_history map<string,string>,

item_info map<string,string>)

ROW FORMAT SERDE 'com.ibm.spss.hive.serde2.xml.XmlSerDe'

WITH SERDEPROPERTIES (

"column.xpath.seller_name"="/listing/seller_info/seller_name/text()",

"column.xpath.seller_rating"="/listing/seller_info/seller_rating/text()",

"column.xpath.bidder_name"="/listing/auction_info/high_bidder/bidder_name/text()",

"column.xpath.location"="/listing/auction_info/location/text()",

"column.xpath.bid_history"="/listing/bid_history/*",

"column.xpath.item_info"="/listing/item_info/*"

)

STORED AS

INPUTFORMAT 'com.ibm.spss.hive.serde2.xml.XmlInputFormat'

OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat'

TBLPROPERTIES (

"xmlinput.start"="<listing>",

"xmlinput.end"="</listing>"

);

3. Load the data into the hive table

load data local inpath '/home/cloudera/ebay.xml' overwrite into table ebay_listing;

4. Display the buying history of CPU.

SELECT seller_name, bidder_name, location, bid_history["highest_bid_amount"], item_info["cpu"] FROM ebay_listing;